2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Towards Better Communication: Refining Hand Pose Estimation in Low-Resolution Sign Language Videos

Sümeyye Meryem Taşyürek*, Tuğçe Kızıltepe*, Hacer Yalim Keles
Department of Computer Engineering, Hacettepe University, Ankara, Türkiye

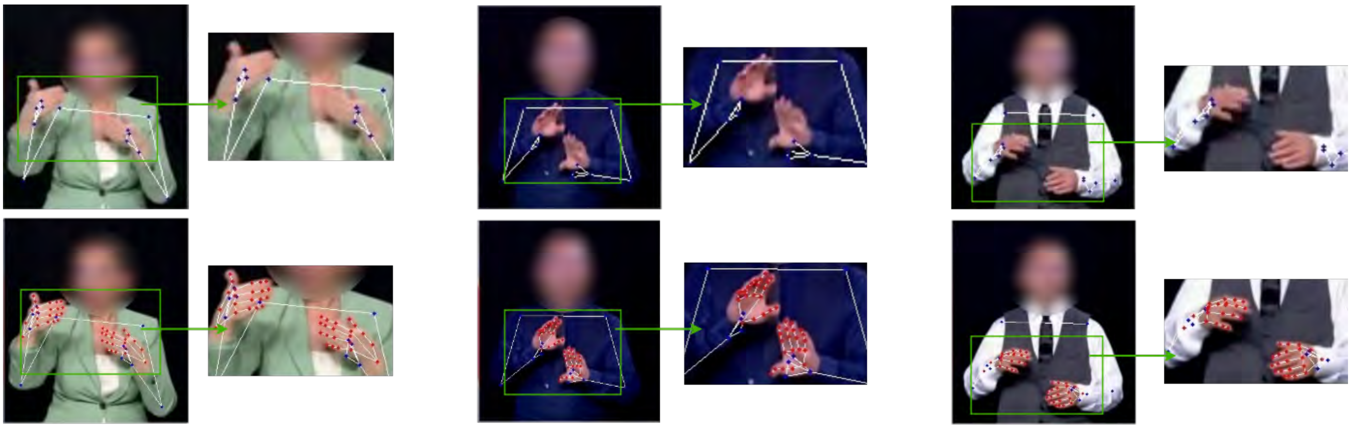

Fig. 1: Sample test frames. Top row: low-resolution frames, bottom row: high-resolution counterparts obtained by our proposed model. Keypoints are superimposed on each frame. All the keypoints are successfully captured in the enhanced high-resolution frames.

***Abstract*— In this study, we present a novel methodology that enhances hand keypoint extraction in low-resolution sign language datasets, a challenge that has been largely unexplored in sign language research. By addressing the limitations of existing pose extraction models like OpenPose and MediaPipe, which frequently struggle with accurately detecting hand keypoints in low-resolution footage, our method marks a notable advancement in this specialized field. Our methodology adapts the U-Net and Attention U-Net architectures to improve the resolution of sign language videos while reducing undetected hand presence (UHP) in low-resolution footage. The key innovation focuses on hand movements through a progressive training procedure, utilizing datasets from SRF DSGS and ShowTV Main News domains. Through comprehensive experimentation and cross-dataset evaluations, our findings demonstrate a significant reduction in the UHP ratio, notably in the Attention U-Net model with our proposed loss function, tailored to enhance hand keypoints detection. In our benchmark tests, using low-resolution TV news broadcasts, our fine-tuned models, particularly the BWA-UNet, showed marked improvements in hand keypoint accuracy compared to standard upsampling methods. These results underscore the effectiveness of our approach in practical, real-world scenarios, highlighting its potential to substantially improve hand keypoint detection in sign language videos.**

I. Introduction

In the key sign language domain tasks, such as sign language production and recognition, there is a preference for utilizing pose sequences rather than the images directly [1], [2], [3], [4], [5]. In these contexts, the accurate extraction of hand keypoints is essentially important for effective communication and interpretation. The manual annotation of keypoints is neither a preferred nor a cost-effective method in the field of sign language. Consequently, pose estimation models are used to generate ground truth and train the models in sign language studies. Despite rapid advancements in this field and the availability of models capable of producing satisfactory outcomes, there remains a significant challenge in achieving accurate hand pose estimation. Current studies in this area often fall short in this regard, highlighting a gap in the field that needs addressing.

Recently published continuous sign language datasets, like Content4All [6] and SRF DSGS Daily News Broadcast [7], make it easier to accurately extract hand keypoints thanks to their high-resolution. However, the practicality of collecting such datasets is limited due to the high costs associated with acquiring a sufficient variety of samples for different sign languages. This challenge is particularly evident when training advanced deep learning models, such as those based on transformers, which require extensive datasets for each distinct sign language. This highlights the need for more accessible and cost-effective data sources. In this context, low-resolution sign language content, commonly found in TV news broadcasts, emerges as an underexploited yet abundant resource. Leveraging this readily available data could significantly enhance the generation of continuous datasets in sign language research, offering a scalable and cost-effective alternative to their high-resolution counterparts.

However, the utilization of low-resolution videos from sources such as TV broadcasts introduces unique challenges for keypoint extraction. Although existing works in the

* Co-first authors

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

super-resolution domain, like the ones proposed by Ledig et al.[8], Lim et al. [9], Chen et al. [10], Zamfir et al. [11] have made significant strides in increasing image resolution, merely enhancing resolution does not directly address the challenge of accurate hand keypoint detection in sign language videos. Motion blur and other related artifacts often persist in the upsampled versions, rendering widely used off-the-shelf pose detectors, such as OpenPose [12] and MediaPipe Holistic [13], ineffective for hand keypoints identification. Our research confronts this issue by focusing on removing such artifacts while increasing the resolution of videos. This approach is pivotal in bridging the gap between enhanced resolution and the effectiveness of pose detectors in interpreting sign language from low-resolution sources.

Our study also contributes to the assessment of model performance. We focus on reducing motion blur and other quality-degrading factors to enable existing pose detectors to accurately capture all hand joints. To validate our approach, we employ a dual evaluation strategy: assessing performance within the existing high quality dataset's test set that we created for our purposes and utilizing a benchmark test set that we derived from low-resolution TV news broadcasts. This, so called 'in the wild' evaluation is a novel aspect of our research, as there is no known precedent in the literature for such an assessment approach. Utilizing datasets such as SRF DSGS Daily News Broadcast, which provide extensive, continuous, high-quality sign language content, as the initial resources in our approach, our objective is to improve the utility of data in scenarios where traditional pose estimation models fail. To achieve this, we have adapted the U-Net architecture [14], known for its effectiveness in super-resolution tasks, introducing an additional upsampling block to increase the resolution of low-resolution images by a factor of two. This study is particularly crucial given the limitations of manual annotation and the reliance on pose estimation models for generating ground truth in sign language studies. We specifically target the enhancement of hand keypoint extraction, acknowledging the difficulties posed by motion blur and other quality-degrading factors in low-resolution videos.

In this context, we also explored various loss functions, moving beyond the traditional Mean Square Error (MSE) approach, which often leads to a loss of high-frequency details and edge blurring in upsampled images. This exploration aims to enhance the recovery of failed detections of hand keypoints from low-resolution videos. We further validate our model's efficacy and generalizability through cross-dataset evaluations using data from ShowTV Main News, a daily broadcast in Turkey. This dataset is particularly relevant for our study, given its low-resolution and motion-blurred content, typical of publicly available sign language videos. Our approach not only addresses the technical challenges of enhancing low-resolution videos for precise keypoint extraction but also underscores the importance of optimizing time and resource allocation in sign language research, demonstrating the potential to significantly advance the field. This advancement is not only academically significant but also holds immense potential for real-world applications, especially in diversifying the data sources for sign language research globally.

II. Related Work

A. Dataset

In the field of sign language research, a diverse array of datasets is available, each with its own unique advantages and limitations. Predominantly, the field features isolated sign language datasets, such as those referenced in [15], [16], [17], [18]. These datasets typically comprise recordings that represent a sign along with its spoken language translation. However, our research is centered on continuous sign language, rendering many high-quality isolated datasets, despite their merits, unsuitable for our specific focus.

The domain of continuous sign language is characterized by a relatively limited number of large-scale datasets that offer broad, non-isolated contexts. A notable example is the Content4All dataset by Camgöz et al. (2021), which encompasses six datasets with a total of 190 hours of video footage, primarily in the news domain. Of these, 20 hours of video with a resolution of 1280x720 pixels have been annotated and made publicly available for research. Another significant dataset is the SRF DSGS Daily News Broadcast, covering daily news from 2014 to 2021 with 30-minute episodes. The dataset for the year 2020 alone includes approximately 60 hours of video footage across 119 episodes, all in 1280x720 resolution. This dataset was notably used in the WMT Shared Task on Sign Language Translation (WMT-SLT22), with the participating teams' experiments and results shared publicly [19]. Given its accessibility, relevance, and recent publication, we have selected this dataset as a primary resource for our study.

Regarding Turkish sign language, there are a few datasets available, but they predominantly consist of isolated sign images [15], [16]. Given our focus on continuous sign language, these datasets do not align with our research objectives. Consequently, we explored publicly available resources and realized the absence of a readily usable dataset in this domain. In attempting to create and utilize our dataset, we encountered challenges with low-resolution and motion blur, which significantly hindered hand detection by pose estimation models. This realization forms the cornerstone of our research, highlighting the need for improved methods to utilize such low-resolution, motion-blurred data effectively.

B. Pose Estimation

Pose estimation is a critical area in computer vision, encompassing a variety of methodologies such as sequential prediction, convolutional architectures, hierarchical models, non-tree models, and classical approaches [20]. Contemporary state-of-the-art research predominantly employs models based on Convolutional Neural Networks (CNN).

A notable example is the DeepPose architecture [21], which utilizes a multi-stage model. Initially, pose estimation is conducted using a DNN-based regression, akin to AlexNet. The outputs of this stage are further refined through a subsequent CNN structure, enhancing the accuracy of the pose estimation.

Another significant approach is embodied by Convolutional Pose Machines [20], a multi-stage system that generates belief maps from the original image in its initial phase. These belief maps, coupled with the original image and an expanding effective receptive field, undergo refinement in subsequent stages, progressively improving the accuracy of the pose estimation.

Earlier studies in pose estimation focused predominantly on single-person scenarios. However, the introduction of part affinity fields [22], a nonparametric representation, marked a significant advancement in multi-person pose estimation. This approach facilitates the association of individual body parts with their corresponding entities in the image. The multi-stage algorithm employed in this study generates a part confidence map, enabling the extraction of body parts without person-specific references. Furthermore, it allows for the effective association of these body parts using part affinity fields.

OpenPose[12] represents a further evolution in multi-person pose estimation, utilizing part affinity fields to perform real-time estimation for multiple individuals. Unlike previous models, OpenPose focuses on refining only the part affinity field throughout its training process. It has become a popular choice for sign language tasks due to its efficiency and accuracy.

MediaPipe Holistic [13] is another multi-stage model extensively used in sign language tasks. It begins by estimating the human pose, followed by generating three distinct crops of interest: two for the hands and one for the face. To enhance the Region of Interest (ROI), a re-crop model is used. This model first extracts the keypoints of the hands and faces using separate, dedicated models before merging them for a comprehensive pose estimation.

Despite the advancements in pose estimation models, challenges persist, particularly in scenarios involving low-resolution and motion-blurred images, where hand detection becomes problematic. This issue is crucial in sign language research, yet it remains inadequately addressed. Camgoz et al. [23] acknowledged this problem, but their study did not propose a solution. Our research aims to address this gap by developing methods to reduce the loss of hand detection accuracy in low-resolution and motion-blurred images, particularly focusing on enhancing the performance of MediaPipe Holistic in such challenging conditions.

C. Super Resolution

The field of super resolution, particularly for single-image enhancement, began with the utilization of traditional techniques like Lanczos resampling [24], bicubic interpolation, and linear filters. While these methods marked initial progress, they often fell short of accurately transferring high-frequency components, leading to suboptimal results. With the advancement of computational techniques, the focus shifted towards more sophisticated methods. Recent studies have demonstrated the effectiveness of CNN-based models, notably the U-Net architecture [25], [26], [27], which has gained prominence due to its proficiency in producing high-quality results, even with limited data.

In the broader scope of computer vision, Generative Adversarial Networks (GANs) have emerged as a powerful tool. A GAN comprises two interconnected networks: the generator, which creates plausible data, and the discriminator, which differentiates between real and generated data. This architecture has been successfully applied in super resolution research, with several studies leveraging GAN models to achieve impressive enhancements in image quality [28], [29], [30].

Additionally, diffusion models, a newer class of generative models, have gained traction in recent years. These models offer an alternative approach to super resolution challenges [31], [32], [33], with some evidence suggesting their potential to outperform GANs in certain applications [34].

Despite the advancements in these complex models, we have chosen U-Net for its efficiency in achieving fast and cost-effective outcomes. It demonstrates strong performance even with limited training data and tends to converge more quickly during training than other architectures. The U-Net architecture, consists of two primary pathways: a contracting path that captures contextual information through convolutional and pooling operations, and an expansive path that allows for precise localization via upsampling and convolutional operations. Integral to its architecture are the skip connections that bridge corresponding layers in the contracting and expansive paths, merging high-level, semantic information from the former with detailed spatial data from the latter.

Several models based on U-Net have been developed for super-resolution. For instance, Hu et al. (2019) introduced a novel U-Net architecture (RUNet) designed to establish correlations between degraded low-resolution (LR) images and their high-resolution (HR) counterparts, incorporating a dynamic degradation model during training for enhanced single-image super-resolution [35]. Another notable adaptation is the Multi-Level U-Net network (MUN) proposed by Han et al. (2022), which employs a multi-level U-Net residual structure. This structure integrates two distinct U-Net frameworks to extract multi-level features from low-resolution (LR) images, offering improved image super-resolution reconstruction [36]. Our choice of U-Net is driven by its adaptability and effectiveness in handling the nuanced requirements of enhancing low-resolution sign language videos for precise keypoint extraction.

Attention U-Net [37] is another U-Net-based adaptation. It employs attention gates to suppress irrelevant information while emphasizing important features. Multiple datasets have demonstrated increased accuracy as compared to the standard U-Net. To improve the accuracy of hand detection in our dataset, we employed Attention U-Net in addition to the standard U-Net model.

D. Loss Functions

A diverse array of loss functions has been devised for effective training in super-resolution modeling. Mean Absolute Error (MAE) and Mean Squared Error (MSE) are commonly used to reduce the discrepancy between predicted and actual ground truth values. Studies have shown that MAE can achieve lower loss values compared to MSE [38]. However, when employed individually, MAE or MSE may cause excessive smoothing or blurring in the super-resolved images. To mitigate this, enhancements are often incorporated into these loss functions, as relying solely on either MAE or MSE is generally not preferred.

A critical challenge in super-resolution is the loss of high-frequency components during the downsampling of images. Preserving these components in the upsampled output is essential. The Mean Squared Canny Error (MSCE) approach, which computes MSE between the edges of the ground truth and predicted images using the Canny edge detector, has been proposed to address this issue [39].

In addition to these pixel-based loss functions, there are perceptual losses that focus on the overall properties of an image [40], [41]. One such metric is the Structural Similarity Index (SSIM), which measures the structural similarity between two images. In the context of super-resolution, employing SSIM loss encourages the generated high-resolution image to maintain structural integrity in comparison to the target high-resolution image [42].

Furthermore, to enhance overall model efficacy, weighted loss functions are strategically designed [43]. To improve our model's accuracy in detecting hands, we implemented a weighted loss function. This function is specifically designed to focus more on important areas, ensuring that the model pays extra attention to the hands. This approach becomes even more effective when combined with super-resolution techniques. It helps the model to concentrate better on hand details, which enhances accuracy in detecting hands.

III. THE METHOD

A. Model Architecture

We utilize the U-Net architecture for our problem, characterized by its distinct U-shaped structure (Fig. 2). This architecture is divided into two primary components: the contracting path, known as the encoder, and the expansive path, referred to as the decoder. The contracting path resembles a traditional CNN architecture with a series of convolutional and max-pooling layers. These layers are responsible for capturing and encoding the high-level features and context of the input image. At the bottom of the U-shape, there is a bottleneck layer that serves as a bridge between the encoder and decoder. It captures the most essential features, acting as a bottleneck to reduce spatial dimensions. The expansive path is the mirror image of the contracting path and is composed of up-sampling and concatenation operations. Up-sampling is used to restore the spatial resolution of the feature maps. Concatenation combines feature maps from the contracting path with those from the up-sampled layers, allowing the model to leverage both high-level context and detailed spatial information. The final layer involves an additional upsampling layer to further increase the spatial resolution two times in each dimension and a convolutional layer with a 1x1 kernel, which reduces the number of channels to match the desired number of channels for the output image.

In our method, we improved the basic U-Net design by adding attention blocks to our architecture, which is depicted in (Fig. 3). Similar to the Attention U-Net architecture [37], we integrated attention blocks at points where encoder skip connections meet the decoder. These blocks filter the the encoder feature maps, using a gating signal from the decoder's current layer that incorporates contextual information from coarser levels. To do that, attention blocks produce attention coefficients, which range between 0 and 1, incorporating coarse features from lower levels and input features from higher levels. The result is a filtered feature map that highlights key features while suppressing the less relevant background. This map is then merged with the decoder's upsampled output, sharpening the model's focus on critical regions.

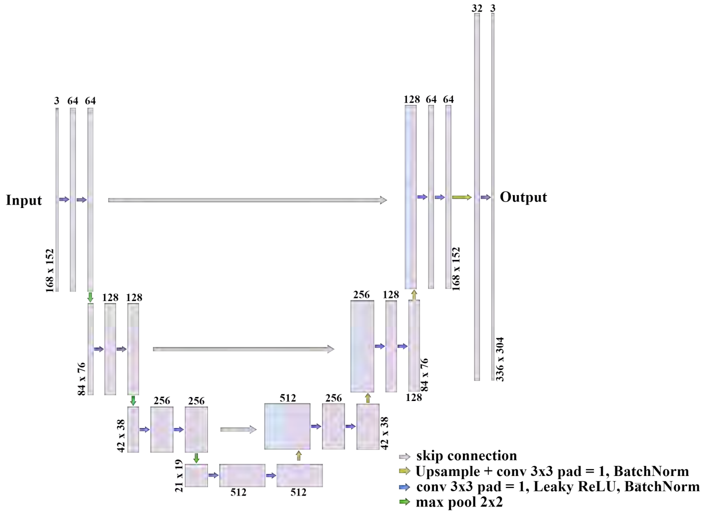

Fig. 2: U-Net architecture

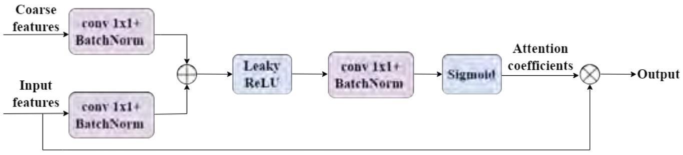

Fig. 3: Block diagram of the Attention Block [37]. It is located at every skip connection, where encoder feature meets the decoder.

B. Datasets

The creation of appropriate datasets was a fundamental aspect of our research, especially in the absence of existing data tailored to our specific objectives. This section elaborates on the extensive and critical process of dataset development, which was both time-consuming and essential for our study. We dedicated significant effort to collecting and organizing data from primary sources, meticulously ensuring that these datasets precisely met the requirements of our analysis. In this section, we detail our methodical approach to selecting and preparing these datasets, highlighting their pivotal role in the validity and success of our research.

We have categorized our datasets into three distinct groups within two domains, the base and target domains, to facilitate a structured approach to our analysis. The first group, designated as *D1*, is from the base domain, comprising the SRF DSGS dataset. This dataset features high-resolution videos, which we have adapted to lower resolutions through a controlled downsampling process. This process enables us to precisely generate both input and target frames, providing a robust foundation for initial model training and adjustments.

Within the target domain, we have divided our data into two subsets from ShowTV Main News. The first subset, referred to as *TD1*, is a smaller dataset carefully prepared for fine-tuning purposes. This subset includes pairs of training and target frames, allowing for precise model refinement. The second subset within the target domain, labeled as *TD2*, serves as our benchmark test set. This test set is crucial for evaluating the model's performance under various conditions and is used exclusively for testing purposes.

The structuring of these datasets into base and target domains, with distinct subsets for training, fine-tuning, and testing, is aligned with our objective to assess model performance from multiple perspectives and for diverse applications.

C. Base Dataset (SRF DSGS)

We selected episodes from the year 2020 of the SRF DSGS Daily News Broadcast dataset as our base dataset in this research. This dataset comprises around 60 hours of video footage, each episode being 30 minutes long and presented in 1280x720@25fps. Notably, the footage exclusively features signers, aligning with our specific requirements.

Given that an extensive amount of data is not essential for training our architecture to enhance resolution, we have strategically utilized 2.5 hours of these videos. We have divided the videos into 1-minute video segments, resulting in a dataset comprising 150 videos, each 1 minute long. Out of these, 120 are allocated for training (i.e., 180,000 frames), while the remaining 30 are for testing (i.e., 45,000 frames).

Additionally, we prepared inputs and ground truths for training our models. To achieve this, we initially cropped the original videos to dimensions of 517x571, ensuring the signer is positioned at the center of the footage. Subsequently, we downsampled the videos to 304x336, aligning with the requirement for inputs to be of size 152x168. This size is necessary for this research, since the low-resolution video frames that we collected from ShowTV Main News have this resolution.

Since the video frames in SRF DSGS dataset are in high quality, we process them to add motion blur to their low-resolution versions. We want to get their distribution as close to the target domain dataset (ShowTV Main News') low resolution data as possible; to enhance the model's ability to handle real-world low-resolution scenarios. To implement this, we used the capabilities of DaVinci Resolve, a professional video editing and color correction software developed by Blackmagic Design. Leveraging this tool, we seamlessly applied artificial motion blur to our videos, carefully adjusting the scale for visual authenticity. Our method involved an initial application of optical flow, followed by the nuanced integration of motion blur across all videos, precisely set at a scale of 3. Example frames can be seen in Fig. 4. The images on the top row show the input images, which are low resolution and processed versions of the original frames, while the bottom row shows the corresponding target images (2 times higher resolution than inputs) used as in the original dataset.

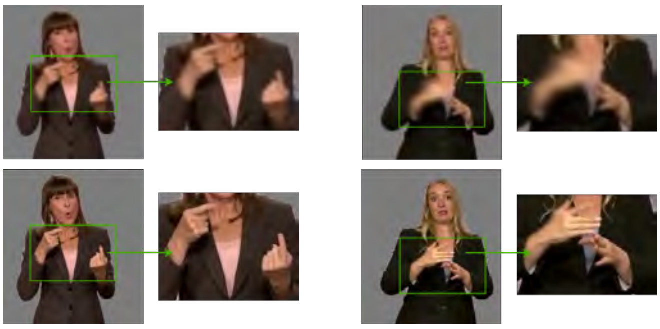

Fig. 4: Examples from the SRF DSGS dataset. Top row: low resolution processed frames (inputs), bottom row: high resolution frames (ground truths).

D. Target Datasets (ShowTV Main News)

A subset of the ShowTV Main News episodes [44], [45], [46], [47] present a unique opportunity due to their stable background and the scarcity of high-resolution data. We utilized a small, carefully selected subset of this data as a benchmark for fine-tuning our models (i.e., 62,000 frames). We will refer to this subset as *TD1*. This fine-tuning process involved choosing episodes with a consistent black background and processing both low-resolution (720p) and high-resolution (1080p) footage from these episodes to generate appropriate inputs and targets (i.e., ground truths). Exemplary frames from this dataset are depicted in Fig. 5.

In addition to the first target set (TD1), which contains input and their ground truths, we created another dataset, namely (TD2) for cross-dataset evaluations of our model from the low-resolution episodes of ShowTV Main News broadcasts, where high-resolution footage was not available in our collected data. Given the realistic nature of this dataset, we do not have access to ground truth data for hand keypoint positions. Therefore, the effectiveness of our models is assessed uniquely by counting the number of frames in which the pose estimation models are able to accurately detect hand keypoints. We will refer to this dataset as Benchmark Test

Set. This approach is critical in our evaluation, as it offers a practical measure of the models' performance in enhancing hand pose estimation, especially in the absence of ground truth benchmarks in these real-world data scenarios.

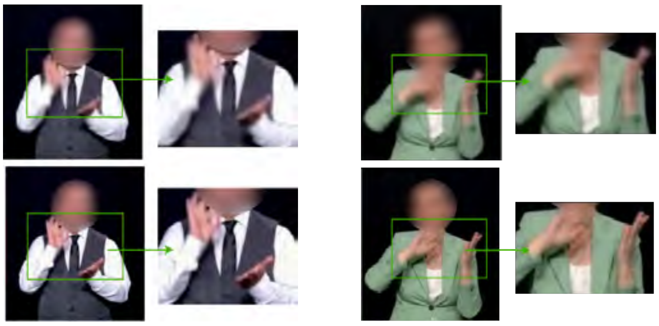

Fig. 5: Fine-tuning data samples from the ShowTV Main News. Top row: low-resolution frames, bottom row: high-resolution frames.

E. Loss Functions

1) SSIM+L1 Loss: In our research, we employed the U-Net model, experimenting with various loss functions to optimize its performance. Specifically, we tested the L1 and L2 loss functions, both individually and in combination with the SSIM loss. Our findings indicated that the L1+SSIM loss configuration yielded the most effective results.

The integration of SSIM with the L1 loss function is particularly advantageous as it targets two critical aspects of image quality. The L1 component focuses on pixel-wise accuracy, ensuring that each pixel in the output closely matches its counterpart in the target image. Meanwhile, SSIM contributes to the perceptual quality of the images, assessing and enhancing the visual similarity between the output and the target. This combination is beneficial because SSIM is inherently less sensitive to minor local distortions. As a result, the combined loss function offers enhanced robustness, effectively handling structural variations in the images. Given these advantages, we adopted the L1+SSIM loss configuration for both the U-Net and the Attention U-Net models in our study.

2) Weighted Loss (WL): Our study focused on enhancing the resolution of input images, specifically targeting the hand regions where detail is often lost due to low resolution and motion-related distortions. While the SSIM+L1 loss function showed promising results for general upsampling, it proved inadequate for refining the hand areas. To address this, we developed a novel loss function, termed *weighted loss*, which emphasizes the hand regions.

This weighted loss function operates by scaling the loss values based on their alignment with a generated map. Utilizing pose information, particularly around the wrist joints of each hand, we generate circular masks centered on these joints. These masks correspond to areas in the generated weight map, which is the same size as the input image, as illustrated in Fig. 6. Within these circular masks, the loss is given twice the weight compared to areas outside the masks.

During the optimization process, the loss for each pixel is adjusted according to its corresponding weight in the map. By doing so, our method ensures more focused attention and improved detail enhancement in the hand regions of the images.

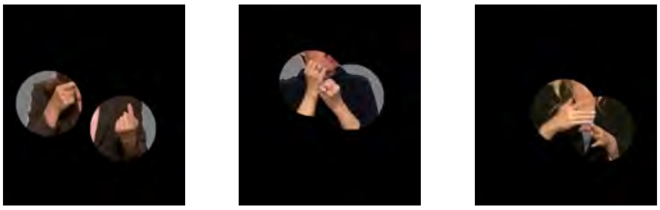

Fig. 6: Hand masks are superimposed on selected frames from sign language videos. These masks are crucial for the weighted loss calculations, emphasizing the hand regions to enhance detail and accuracy in the model's output.

IV. EXPERIMENTS

A. Training Procedure

The training of our models was conducted on a computer equipped with an Nvidia RTX A4000 GPU, utilizing Pytorch framework. We consistently used the Adam optimizer with its default parameters, setting the learning rate at 0.001. Each training session was configured to run for 50 epochs, but we incorporated an early termination feature: if the validation loss did not improve for 7 consecutive epochs, the training was automatically halted.

We have a progressive training procedure with the following stages:

Initial training: In this stage, we train models using the Base Dataset (*D1*) and then evaluate their performance on both *D1* and the benchmark dataset (*TD2*). We removed the skip connections from the U-Net architecture and transformed it into an autoencoder model referred to as B-AE in order to investigate how the skip connections in the U-Net model improve performance. After that, the standard U-Net and Attention U-Net architectures are employed, utilizing SSIM+L1 loss on *D1*, resulting in the B-UNet and BA-UNet models, respectively. Both models undergo evaluation using the test sets of *D1* and *TD2*, facilitating both within-domain and cross-domain assessments. Here, the BA-UNet model exhibited higher performance.

Model Enhancement with Weighted Loss: Building on the results of initial training, we introduced another model, trained using *D1* with our proposed Weighted Loss (WL) in the Attention U-Net architecture, creating the BWA-UNet model. This model incorporates a weighted SSIM+L1 loss, with double weight assigned to the hand region. The

BWA-UNet showed superior performance on the *TD2*.

Fine-Tuning with Target Dataset: In this stage, the best-performing model from the earlier stages is fine-tuned using the TD1 dataset. We denote this model as BWFA-UNet. This model is also evaluated against *TD2* to assess improvements.

B. Evaluation Procedure

Our evaluation procedure is designed to assess model performance through a video-based approach, contrasting the frame-by-frame analysis during the training phase. During testing, we process each frame of the test dataset videos, each lasting one minute and comprising approximately 1500 frames. The models are provided with these frames, and the resulting outputs are then composed into one-minute video segments. These segments are fed into Mediapipe for pose extraction. Mediapipe Holistic model is used with the parameters, min_detection_confidence=0.3, min_tracking_confidence=0.3.

The key metric for our evaluation focuses on the detection accuracy of the dominant hand (the right hand in our datasets). We utilize an automated process to extract data from frames where the dominant hand is not successfully identified by the model. This data enables us to calculate the ratio of frames with *undetected hand presence (UHP)* to the total frame count. The average ratio derived from various test videos serves as the primary measure for assessing the efficacy of our model for consistent hand pose detection across continuous video footage.

Additionally, we consider key metrics: Peak Signal-to-Noise Ratio (PSNR) measures the quality of the reconstructed image; Structural Similarity Index (SSIM) evaluates similarity between original and reconstructed images; Perceptual Hashing (pHash) generates a compact image representation with a 64-bit hash value; Hamming distance measures similarity between two images based on differing bits; Fréchet Inception Distance (FID) quantifies similarity between the distributions of real and generated images, assessing quality and diversity.

C. Results and Discussion

In order to assess our models, we first generated upsampled videos using standart upsampling function with bilinear interpolation. Following our training procedure, we trained the baseline models that we planned for the initial stage, namely B-UNet, BA-UNet and BWA-UNet on base dataset (D1). The performances are depicted in Table I.

Before going into the details of the models, let us explain the average UHP Frame numbers and their ratios for the input videos and the target (2x size) videos in the D1 dataset; in 33.13% of the input videos, the dominant hand cannot be identified by the Mediapipe. On the other hand, in the target videos (ground truths), only 2.47% are lost. This shows that the prepared dataset could be effectively utilized for training a deep model to reduce UHP.

The UHP ratio of the standard bilinear upsampling algorithm is worse than the UHP ratio of the original input videos (by around 2.97%). This outcome clearly demonstrates that simply upscaling inputs using conventional methods is inadequate for maintaining, let alone enhancing, hand detection accuracy. It underscores the importance of developing a learned approach to upsampling that is specifically tailored to preserve and accentuate critical features, such as hand poses, in the upscaled output. This is essential for ensuring that key details are not lost or distorted in the process of increasing resolution.

Among our baseline models, the BWA-UNet emerges as the top performer, notably enhancing the UHP ratio compared to the original input videos. Specifically, the BWA-UNet reduces the number of frames with lost hand keypoints from 497 to 99, translating to a significant decrease in loss to 6.60%. When comparing B-AE with B-UNet, B-UNet demonstrated a 0.27% enhancement over D1 and a 1.13% improvement over TD2. This demonstrates that incorporating skip connections into a model improves its performance. However, when compared to the BA-UNet model (Attention U-Net without weighted loss), the improvement is marginal. This observation suggests that the addition of weighted loss to the Attention U-Net model results in only a slight performance advantage in recovering hand keypoints from low-resolution videos. The performance of BWA-UNet and BA-UNet is notably similar, indicating that the weighted loss does not significantly enhance the model's effectiveness as initially expected. Sample results of the BWA-Unet model are shared in Fig. 7.

BWA-UNet also outperformed other models based on several performance metrics except FID score where BA-UNet performed slightly better than BWA-UNet. A higher Peak Signal-to-Noise Ratio (PSNR) number implies a lower level of distortion or noise. Similarly, when comparing the structural similarity of two images using the Structural Similarity Index (SSIM), a higher value implies a greater degree of similarity. A phash value near 0 implies a significant amount of similarity between the images. Consistent with the results we obtained according to the UHP ratio metric, BWA-UNet produced more successful results in these metrics. Upon examining the results, it is noticeable that bilinear interpolation is not effective. Nevertheless, it can be stated that UNet models produce very similar results.

Given our objective of reducing the occurrence of undetected hands in frames, we cannot solely rely on visual enhancements when utilizing pose estimation models. Therefore, while performance metrics like SSIM, PSNR, and phash indicate an improvement in quality, they do not offer specific insights into eliminating the occurrence of undetected hands. Therefore, the UHP ratio is an essential metric for us.

After evaluating our baseline models, we proceeded to conduct cross-dataset evaluations to further assess model performance. This involved using the base models initially trained on D1, and then evaluating them both with and without fine-tuning on TD1, against TD2 (our benchmark test dataset). Sample frames from TD2 are depicted in Fig. 1, together with the results we obtained with our best performing model (i.e., BWFA-UNet). UHP ratios on TD2

TABLE I: Evaluation Results in the D1 (SRF DSGS) Test Set

Model	Loss	UHP-Frame#	UHP-Ratio(%)	PSNR	SSIM	pHash	FID
Input Videos	-	497	33.13	-	-	-	451.786
Target Videos	-	37	2.47	-	-	-	-
2x Bilinear Interpolation	-	540	36.00	29.33	0.891	0.752	467.083
B-AE	SSIM+L1	110	7.33	33.893	0.948	0.831	64.202
B-UNet	SSIM+L1	106	7.06	34.588	0.951	0.574	53.633
BA-UNet	SSIM+L1	100	6.67	34.576	0.952	0.571	**49.884**
BWA-UNet	WL	**99**	**6.60**	**34.610**	**0.953**	**0.536**	50.539

TABLE II: Evaluation Results in the TD2 (ShowTV) Test Set

Model	Loss	UHP-Frame#	UHP-Ratio(%)
Input Videos	-	390	26.00
2x Bilinear Interpolation	-	401	26.73
B-AE	SSIM+L1	369	24.6
B-UNet	SSIM+L1	352	23.47
BA-UNet	SSIM+L1	338	22.53
BWA-UNet	WL	330	22.00
BWFA-UNet	WL	313	20.87

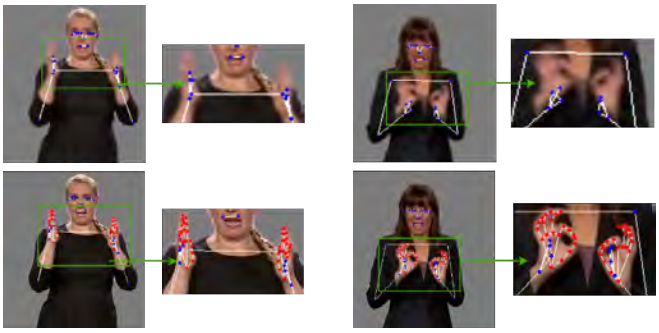

Fig. 7: Detected Mediapipe keypoints; top row: sample inputs of the D1 test set, bottom row: corresponding BWA-UNet model outputs.

for pre-trained and fine-tuned models are presented in Table II.

As is seen from Table II, without any fine-tuning, the best performing model is again the BWA-UNet model. We selected this pretrained model (on D1) and fine-tuned it using TD1 (ShowTV dataset with target images) to observe the improvements of this adaptation to the related domain (from SRF DSGS to ShowTV Main News). Fine-tuning helps improving the results by around 1.13%; the fine-tuned BWA-UNet model exhibited better performance than the other models. This was particularly evident in the cross-dataset evaluations, highlighting the effectiveness of the BWA-UNet model when fine-tuned with TD1, in accurately performing on TD2. The enhancement degree is evident in the sample frames' hand keypoints in Fig. 1.

V. CONCLUSION AND FUTURE WORKS

Our goal was to improve the resolution of low-resolution sign language data, facilitating better hand detection by Mediapipe Holistic in frames where it initially failed. To this end, we integrated an additional upsampling layer into the U-Net architecture, aiming to enhance frame resolution and visual quality. We trained the standard U-Net model, first without attention blocks, and explored a mix of traditional loss functions and the Structural Similarity Index (SSIM) for optimal results. Our initial findings indicated a reduction in failed hand detection in the base dataset, but there was room for further improvement.

To advance our approach, we enhanced the standard U-Net with attention blocks and implemented a weighted loss strategy, prioritizing the hand regions in the images, resulting in the BWA-UNet model. This model nearly reached the minimum ratio of failed hand detections in the base dataset. However, its performance on the TD2 dataset was not as effective, likely due to the differences in data distribution between the two datasets. To improve its generalization capability, we fine-tuned the BWA-UNet model using the TD1 dataset, leading to the development of the BWFA-UNet model, which showed improved performance on the TD2 dataset.

This study represents an initial direction in the field of sign language video processing, particularly in enhancing keypoint detection in low-resolution frames. The advancements demonstrated in our work hold significant potential for the creation of high-quality datasets, especially in scenarios with limited resources. This aspect is crucial, as it paves the way for more accessible and cost-effective methods in dataset generation, enabling broader research and application opportunities in sign language recognition and analysis with minimal financial constraints. The promising results achieved in this study suggest ample scope for further refinement and innovation in this domain, potentially leading to substantial improvements in the accuracy and reliability of keypoint detection in low-resolution sign language videos.

References

[1] S. Stoll, N. Camgoz, S. Hadfield, and R. Bowden, "Sign language production using neural machine translation and generative adversarial networks," 08 2018.

[2] B. Saunders, N. C. Camgöz, and R. Bowden, "Progressive transformers for end-to-end sign language production," *CoRR*, vol. abs/2004.14874, 2020.

[3] B. Saunders, N. C. Camgöz, and R. Bowden, "Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks," *CoRR*, vol. abs/2103.06982, 2021.

[4] B. Saunders, N. C. Camgoz, and R. Bowden, "Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production," 2022.

[5] P. Selvaraj, G. N. C., P. Kumar, and M. M. Khapra, "Openhands: Making sign language recognition accessible with pose-based pretrained models across languages," *CoRR*, vol. abs/2110.05877, 2021.

[6] N. C. Camgöz, B. Saunders, G. Rochette, M. Giovanelli, G. Inches, R. Nachtrab-Ribback, and R. Bowden, "Content4all open research sign language translation datasets," *CoRR*, vol. abs/2105.02351, 2021.

[7] Z. Jiang, M. Müller, S. Ebling, A. Moryossef, and R. Ribback, "Srf dsgs daily news broadcast: video and original subtitle data (version 1.0.0)," *LaRS - Language Repository of Switzerland*, 2023.

[8] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016.

[9] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," *CoRR*, vol. abs/1707.02921, 2017.

[10] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," 2023.

[11] E. Zamfir, M. V. Conde, and R. Timofte, "Towards real-time 4k image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[12] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," 2019.

[13] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," *CoRR*, vol. abs/1906.08172, 2019.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, 2015.

[15] O. M. Sincan and H. Y. Keles, "AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods," *CoRR*, vol. abs/2008.00932, 2020.

[16] O. Özdemir, A. A. Kindiroglu, N. C. Camgöz, and L. Akarun, "Bosphorussign22k sign language recognition dataset," *CoRR*, vol. abs/2004.01283, 2020.

[17] S. Ebling, N. C. Camgöz, P. Boyes Braem, K. Tissi, S. Sidler-Miserez, S. Stoll, S. Hadfield, T. Haug, R. Bowden, S. Tornay, M. Razavi, and M. Magimai-Doss, "SMILE Swiss German Sign Language Dataset," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), European Language Resources Association, May 2018.

[18] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," pp. 1221–1230, 10 2017.

[19] M. Müller, S. Ebling, E. Avramidis, A. Battisti, M. Berger, R. Bowden, A. Braffort, N. Cihan Camgöz, C. España-bonet, R. Grundkiewicz, Z. Jiang, O. Koller, A. Moryossef, R. Perrollaz, S. Reinhard, A. Rios, D. Shterionov, S. Sidler-miserez, and K. Tissi, "Findings of the first WMT shared task on sign language translation (WMT-SLT22)," pp. 744–772, Dec. 2022.

[20] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *CoRR*, vol. abs/1602.00134, 2016.

[21] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, June 2014.

[22] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *CoRR*, vol. abs/1611.08050, 2016.

[23] A. Moryossef, I. Tsochantaridis, J. Dinn, N. C. Camgöz, R. Bowden, T. Jiang, A. Rios, M. Müller, and S. Ebling, "Evaluating the immediate applicability of pose estimation for sign language recognition," 2021.

[24] C. Duchon, "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology - J APPL METEOROL*, vol. 18, pp. 1016–1022, 08 1979.

[25] A. Banerjee and A. Palrecha, "Mxr-u-nets for real time hyperspectral reconstruction," 2020.

[26] S. Peng, C. Guo, X. Wu, and L.-J. Deng, "U2net: A general framework with spatial-spectral-integrated double u-net for image fusion," in *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, ACM, Oct. 2023.

[27] U. Kumaravelan and N. M, "Localized super resolution for foreground images using u-net and MR-CNN," *CoRR*, vol. abs/2110.14413, 2021.

[28] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "Esrgan: Enhanced super-resolution generative adversarial networks," 2018.

[29] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," *CoRR*, vol. abs/1909.06581, 2019.

[30] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "Ranksrgan: Generative adversarial networks with ranker for image super-resolution," *CoRR*, vol. abs/1908.06382, 2019.

[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, vol. abs/2112.10752, 2021.

[32] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," 2023.

[33] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, "Implicit diffusion models for continuous super-resolution," 2023.

[34] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *CoRR*, vol. abs/2105.05233, 2021.

[35] X. Hu, M. A. Naiel, A. Wong, M. Lamm, and P. Fieguth, "Runet: A robust unet architecture for image super-resolution," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 505–507, 2019.

[36] N. Han, L. Zhou, Z. Xie, J. Zheng, and L. Zhang, "Multi-level u-net network for image super-resolution reconstruction," *Displays*, vol. 73, p. 102192, 2022.

[37] O. Oktay, J. Schlemper, L. L. Folgoc, M. C. H. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *CoRR*, vol. abs/1804.03999, 2018.

[38] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "On mean absolute error for deep neural network based vector-to-vector regression," *IEEE Signal Processing Letters*, vol. 27, p. 1485–1489, 2020.

[39] R. K. Pandey, N. Saha, S. Karmakar, and A. G. Ramakrishnan, "MSCE: an edge preserving robust loss function for improving super-resolution algorithms," *CoRR*, vol. abs/1809.00961, 2018.

[40] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016.

[41] S. D. Sims, "Frequency domain-based perceptual loss for super resolution," 2020.

[42] D. J. B. Nay, "Single image super resolution using espcn – with ssim loss," *SSRN*, 2021.

[43] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *CoRR*, vol. abs/1708.02002, 2017.

[44] "Show ana haber - 01 Şubat 2022." https://www.showtv.com.tr/programlar/videolar/show-ana-haber-01022022/102502. [Accessed: March 23, 2024].

[45] "Show ana haber - 1 haziran 2022." https://www.showtv.com.tr/programlar/videolar/show-ana-haber-01062022/105277. [Accessed: March 23, 2024].

[46] "Show ana haber - 2 nisan 2022." https://www.showtv.com.tr/programlar/videolar/show-ana-haber-02042022/103900. [Accessed: March 23, 2024].

[47] "Show ana haber - 19 ocak 2022." https://www.showtv.com.tr/programlar/videolar/show-ana-haber-19012022/102187. [Accessed: March 23, 2024].

[48] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *CoRR*, vol. abs/1807.10165, 2018.

[49] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” *CoRR*, vol. abs/2102.04306, 2021.

[50] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *CoRR*, vol. abs/1501.00092, 2015.

[51] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, “Loss functions for image restoration with neural networks,” *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2017.

[52] Y. Sun, A. Lu, and L. Yu, “Weighted-to-spherically-uniform quality evaluation for omnidirectional video,” *IEEE Signal Processing Letters*, vol. 24, pp. 1408–1412, 2017.

[53] Y. Zhou, M. Yu, H. Ma, H. Shao, and G. Jiang, “Weighted-to-spherically-uniform ssim objective quality evaluation for panoramic video,” pp. 54–57, 2018.